VFAST Transactions on Software Engineering
http://vfast.org/journals/index.php/VTSE@ 2017, ISSN(e): 2309-6519; ISSN(p): 2411-6327
Volume 12, Number 1, January-April , 2017 pp. 10–18

EFFICIENT PREDICTION OF LIVER DISEASE USING SELECTED ATTRIBUTES

MUJTABA HASSAN[1], MAHAM IRFAN[1], SALAh-U-DIN AYUBI [1]

Department of Software Engineering, University of Management and Technology, Punjab, Pakistan

Email: mujtabasheikh555@gmail.com

ABSTRACT. *Liver plays a vital role in the human body that performs several crucial life functions. A number of liver diseases exist and it is a challenging task to diagnose the liver disease at its early stage. In recent years, several data mining techniques have been used in medical field for prediction but there can be further improvements for quick and accurate diagnose of liver disease. In this paper, a variety of Classifiers have been experimented on Indian liver disease patients dataset which is publicly available on Kaggle. Attribute subset selection is performed to identify significant attributes and the resulting dataset is named as Selected Attributes Dataset (SAD). SAD provides more accuracy in less computation time using Random forest classification algorithm and improved system including these parameters i.e., the efficiency of the system can be increased, early decision making, less time and space required. This research work will provide help to predict liver disease with less amount of data, i.e., number of attributes.*

Keywords*:* Data mining, liver disease prediction, classification techniques

1. **Introduction.** Data mining is an activity of finding hidden knowledge and useful pattern from large datasets, warehouse and other repository. As the medical industry collects an immense amount of data. In medical environment we still have rich information and weak knowledge. Data mining is widely used in the medical field for predicting the disease from data produced on a daily basis from patients, diseases, hospital resources, diagnose methods and electronic records. By mining into these data hidden pattern and relationship can be discovered for efficient analysis, diagnoses and prognosis. Medical data mining predicting many diseases, but liver disease is still very challenging to predict in early stages. The liver is the largest internal organ in the human body that plays an important role in many body functions like protein production and blood clotting, cholesterol, glucose, and iron metabolism. There are hundred different forms of liver disease that affecting men, woman and children. Different viruses like; hepatitis A, hepatitis B, hepatitis C, drinking too much alcohol, drugs and poisons. Persons with liver disease has some more symptoms they are dark urine, pale stool,easy bleeding, itching, spider like blood vessel visible in the skin, enlarged spleen, fluid in abnormal cavity, chills pain from the biliary track or pancreas, enlarged gallbladder, impaired brain function and general failing health. Predicting the liver disease at an early stage is highly challengeable for doctors as it will be functioning normally even when it is partially damaged. Different techniques have been proposed by different authors for the prediction of liver disease. Karthik et.al **[1]** used a soft computing technique for diagnosis of liver disease. He has implemented classification and its type detection in three phases. Firstly, classified liver disease using artificial neural network (ANN) classification algorithm. Secondly, generated the classification rule induction using (LEM) and in the third phase fuzzy rules were applied to identify liver disease. Sindhuja et .al, **[2]** described the advantages and disadvantages of algorithms such as C4.5, Naive Bayes, Decision Tree, SVM, Back propagation and Classification and Regression trees are compared. It concludes that C4.5 gives better performance than other algorithms. Ratnamala Kiruba. H, et. al **[3],** combined two different types of liver

disease disorders datasets. Auhtor used C4.5 decision tree algorithm and Random tree algorithm for the prediction of liver disease. Further concludes that both algorithms give good accuracy for predicting the liver disease. Many researchers have done good work related to prediction of liver disease and used different data mining techniques on liver disease patients data. Different authors had applied different data mining techniques on Indian liver disease patients dataset which is publicly available on Kaggle and shown their comparison results for which algorithm¥technique gives better accuracy to predict liver disease. But the problem is they have applied different data mining techniques by using less dataset with less attributes of Indian liver disease patient data (e.g. **29** datasets with **12** different attributes and **345** instances with **7** different attributes) and shows higher accuracies of algorithms. When we studied their research work and saw those results, then we have decided to take a whole dataset included **583** instances with **11** attributes of Indian liver disease patients as available on kaggle (UCI Machine Learning Repository). We have applied different data mining techniques including data preprocessing and in preprocessing, did data reduction and data cleaning. After that Attribute subset selection is performed to identify significant attributes and the resulting dataset is named as Selected Attributes Dataset (SAD). SAD provides more accuracy in less computation time using Random forest classification algorithm and improved system including these parameters i.e., the efficiency of the system can be increased, early decision making, less time and space required.

2. Related Work. Bendi Venkata Ramana et. al **[4]** said AP liver dataset is better than the UCLA liver dataset. Because by using classification algorithms they support the vector machine, C4.5 and Naive Bayes classifier. Aneeshkumar. A. S, et. al **[5]** took 15 attributes of real medical datasets and use classification techniques to classify data of liver and non-liver disease. After classification of data did data cleaning. Divide the data set into three categories ratio based on average and standard deviation of each factor and then evaluate accuracy. C4.5 and Naive Bayes algorithms applied on the datasets and found that C4.5 algorithm gives better accuracy than a Naive Bayes algorithm. Ratnamala Kiruba. H, et. al **[3]** used C4.5 decision tree algorithm and Random tree algorithm to predict. He combined two different types of datasets and predicts the accuracy. These both algorithms give good accuracy to predict liver disease. Dhamodharn. S, et. al **[6]** compared two decision tree algorithms that are FT growth and Naive Bayes and found out which algorithm gives better accuracy. He found that Naive Bayes gives (75.54%) more accuracy than FT growth algorithm (72.66%) using WEKA tool. Rajeswari. P, et. al **[7]**, describes the blood test taken when a person is affected with liver disorder such as alkaline phosphatase, alanine aminotransferase, aspartate aminotransferase, gamma-glut amyl and transpeptidase. According to the attributes the dataset is divided into two parts that is 70% of the data are used for training and 30% are used for testing. Different algorithms used to predict the accuracy of liver disease and these algorithms selected on their performances at the time of training set not around of data set. These three different supervised machine learning algorithms derived from the WEKA data mining tool which includes: Naive Bayes, KStar, and FT Tree. WEKA tool is used to classify the data and the data is evaluated using 10-fold cross validation and the results are compared. Banu, MA Nishara et. Al **[8],** defined that the successful applications of data mining are running in different industries i.e. Ecommerce, Trade. The medical environment has a great amount of information, but less knowledge to identify relationships and trend between data, there is a lack of powerful tool. An approach which is used by the authors is; firstly does preprocessing then they apply k-mean clustering on the dataset. Cluster relevant data and then apply maximal frequent item set algorithm. When frequent patterns have been selected, classifies the pattern using C4.5 algorithm and finally display accuracy and effective heart attack level. Alfisahrin, Sadiyah et. al **[9],** took 10 important attributes of liver disease using Naive Bayes, Decision Tree and NB Tree algorithms and described that the NB Tree algorithm has the highest accuracy, while the Naive Bayes algorithm gives the fastest computation time. They made models of all the algorithms and calculate their accuracy. A confusion matrix of each algorithm is drawn and confusion matrix shows details of misclassification, classification accuracy is a measure that indicates how well a classifier to identify objects correctly. WEKA tool is used to make the model and calculate accuracy. Newton Cheung et. al **[10]**, using data mining classification techniques he found various results using; C4.5 algorithm gives 65.59%, Naive Bayes gives 63.39%, BNND (Bayesian Network with Naive Dependence) gives 61.83% and BNNF (Bayesian Network with Naïve Dependence & Feature Selection) gives 61.42%. The study examines different algorithms such as C4.5, Naive Bayes, Decision Tree, Support vector machine, Back propagation neural network and Classification and Regression Tree algorithms. These algorithms give different result based on speed, accuracy, performance and cost. It is shown that C4.5 gives better results as compared to other algorithms.

Table-1: Work done by different authors in the medical field

Authors	Data Mining Techniques	Dataset	Results/Accuracy
Dhamodharn. S (2014)	FT growth and Naive Bayes	29 datasets with 12 different attributes	Naive Bayes (75.54%) FT growth algorithm (72.66%)
Aneeshkumar. A. S (2012)	C4.5 and Naive Bayes	15 attributes of real medical data	C4.5 gives better accuracy than Naive Bayes
P. Rajeswari and G. Sophia Reena (2010)	Naive Bayes, KStar and FT Tree	345 instances with 7 different attributes	Naive Bayes gives (96.52%) FT Tree gives (97.10%) KStar gives 83.47%

3. Classification Algorithms

A. Support Vector Machine (SVM). A support Vector Machine separates the data into two categories of performing classification and constructing an N-dimensional hyper plane. These models are closely related to neural networks. This model uses a sigmoid kernel function which is equivalent to a two-layer perception neural network. In SVM a transformed attribute is used to define the hyper plane, which is called a feature. A set of feature that describes one case is called a vector. The goal is to design a hyper plane that classifies all training vectors in two classes in such a way that cases with one category of the target variable are on one side of the plane and cases with other category are on the other side of the plane. The vector near the hyper plane is the support vector.

B. Multilayer perceptron. A multilayer perceptron is a class of the feed forward artificial network. An MLP consists of at least three layers of nodes. Except for the input nodes, each node is a neuron that uses a nonlinear activation function. MLP utilizes a supervised learning technique called back propagation for training. Its multiple layers and non-linear activation distinguish MLP from a linear perceptron. It can distinguish data that is not linearly separable.

C. Random Forest. Random forest is a supervised classification algorithm. Build up multiple trees using algorithms such as: information gain, GIGI index approach and other decision tree algorithms. It can be used for both classification and regression. It can Handle the missing values and maintain accuracy of missing data. When you have more trees, random forest will not over fit. It can handle large data set, high dimensionality.

D. Bayesian Logistic Regression In Logistic regression the dependent variable of the regression is a logistic function and the dependent variable is categorical. It is one of the widely used models in problems where the response is a binary variable, for example (fraud or not-fraud, click or no-click) and so on. Logistic regression is used in various fields, including machine learning, most medical fields, and social sciences.

4. Methodology/Approach Used. We took whole dataset of Indian liver disease patients same as available on kaggle (UCI Machine Learning Repository). Some missing values exist in a dataset that can be cleaned by using data mining techniques. We have applied data mining rules on the dataset and preprocessed the data first. In preprocessing, we filled the missing values by taking Mean because values are numeric and then apply different classification techniques on the whole dataset containing "**583** instances with **11** attributes". Different results came after applying different classification techniques like; Bayesian Logistic Regression, Multilayer Perceptron, SVM, Attribute Selected Classifier, Classification via Regression, NBTree, J48 and

Random Forest these all algorithms give different accuracies for prediction of liver disease. In this case, Bayesian Logistic Regression and SMO/SVM give better accuracy than others, but Bayesian Logistic Regression has less computation time too. After getting these results we reduce the dimensionality of data by selecting most relevant attributes using Decision tree and ID3/C4.5 algorithms on the bases of high information gain. For the confirmation of attributes selection used two methods, one is attribute evaluator i.e., CfsSubset Eval has been used along with search method "Best First" and the second is attribute evaluator i.e., CfsSubset Eval used along with search method "Greedy Stepwise" and both methods select the same attributes

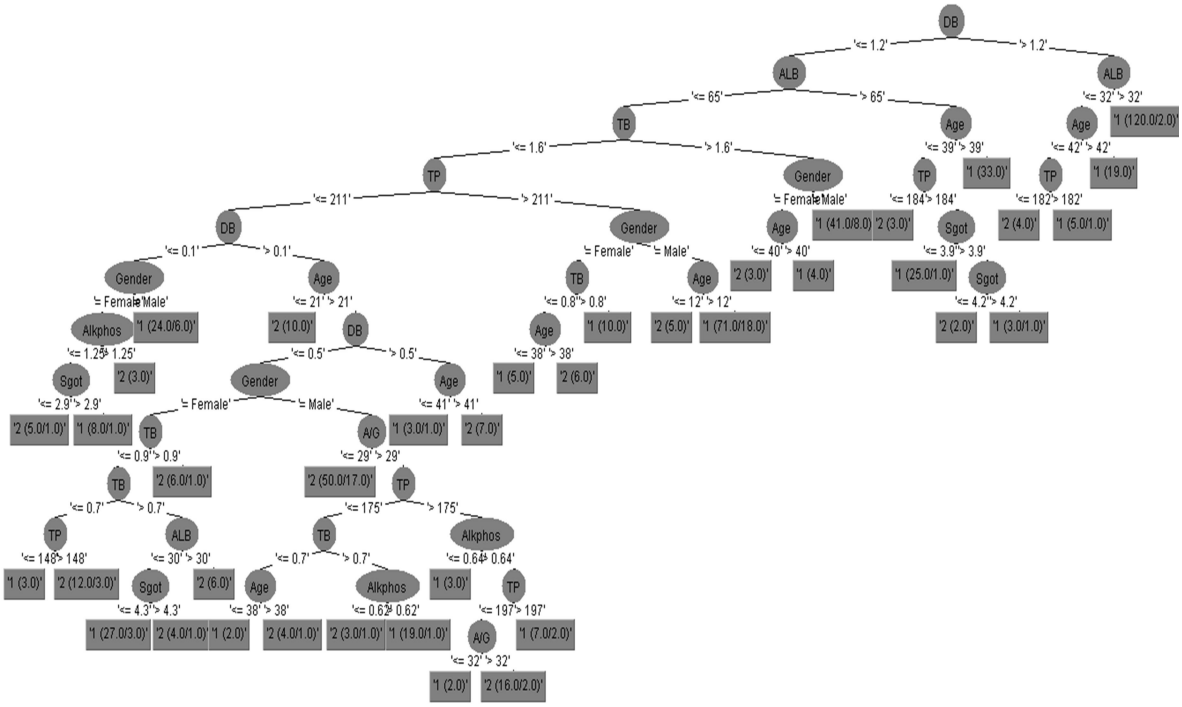

Figure-1: Tree View (Decision tree for Whole Attribute Dataset(WAD))

A. Decision Tree. Decision tree builds classification or regression models in the form of a tree. A decision tree is a structure that includes a root node, branches and leaf nodes. Each internal node denotes a test on an attribute, each branch denotes the outcome of a test and each leaf node holds a class label. The top most nodes in the tree are the root node. The final result is a decision tree with nodes and leaf nodes. We have used two ways; **One** is Whole Attribute Dataset and **Second** is Selected Attribute Dataset, for the prediction of liver disease with Decision tree algorithm. So, we analyze that which way is more suitable to predict liver disease.

We took Indian liver disease patients whole dataset with all attributes and apply decision tree algorithm on it.

You can see the Decision tree for whole attribute dataset above in Figure 1.0. This tree is big and take too much time to take decision and required maximum space

Here in the Figure 2, after applying Decision tree algorithm on both type of datasets which have been prepared by us. We got different trees on the bases of high information gain. From these results we can analyse that less number of attributes dataset gives more accurate results for the prediction of liver disease. It provides ease to get more accurate results as well as improved system including these parameters i.e., the efficiency of the system can be increased, early decision making, less time and space required.

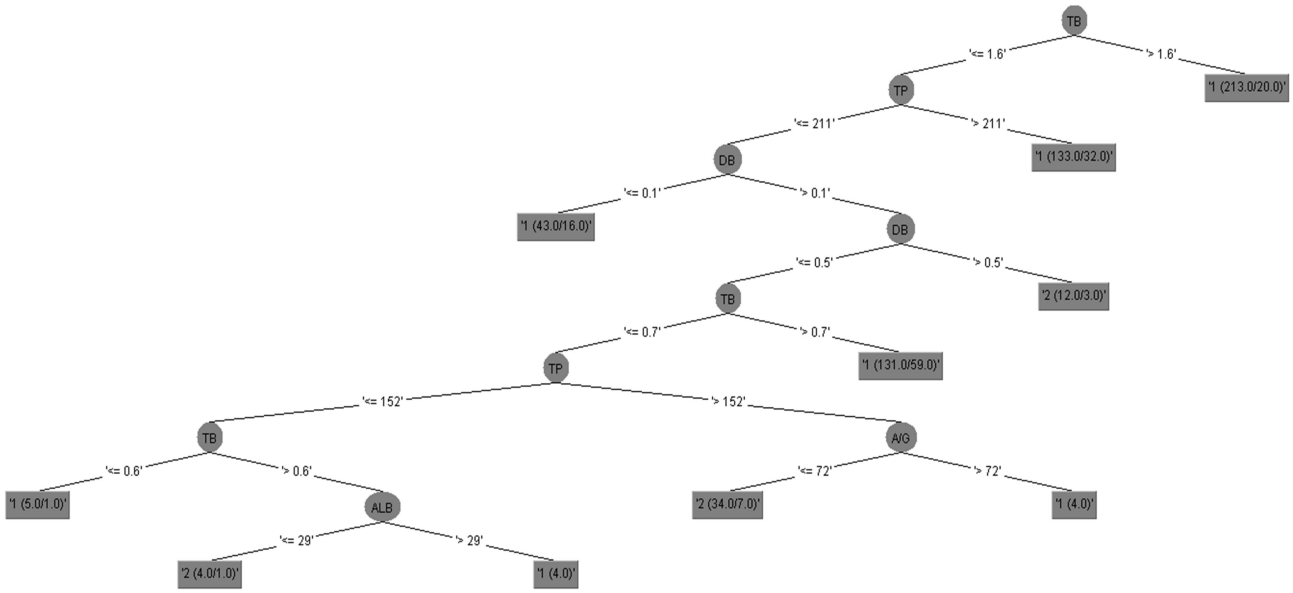

Figure 2 Tree View (Decision tree for Selected Attribute Dataset (SAD))

5, Results And Analysis. Experimental setup, evaluation methodology and evaluation measures followed by results and analysis are discussed in this portion.

A. Experimental Setup. The experimental setup used in our work consists of Microsoft Excel and WEKA.

a. Microsoft Excel. Microsoft Excel is commonly used spreadsheet. It contains feature for graphing, calculations and tables, etc.

We used excel for the following purposes:

- To populate our whole dataset.
- Remove additional attributes and populate selective attribute dataset.
- To fill missing values of attributes by taking median

B. WEKA. WEKA is a tool for data mining that containing machine learning algorithms. It is a collection of different tools for classification, data pre-processing, attribute selection and clustering etc. We have used it for our evaluation experiments and to calculate the accuracy of machine prediction. We have used 10 folds for cross validation along with whole attribute dataset and selective attribute dataset. For example, if we have 100 records, we will create 10 folds for this i.e., Creating 10 sections, each containing 10 files. The number of files, the less number of folds is created. We perform training on more dataset and testing is performed on a small dataset. We use one third part of our dataset for testing because more training will produce better results and performance. For instance, if we have 10 folds, 1 fold will be used for testing and remaining 9 folds will be used for training. Testing will be performed on each fold one by one by replacing the training folds with testing folds. Cross validation gives better results because of repeated testing. We have used WEKA for attribute selection by using attribute evaluators and search methods on the whole dataset. For this purpose first we load dataset in WEKA, after data preprocessing we select Attribute Evaluator and then select the Search method and apply it on the dataset. WEKA return result by giving lists of selected attribute. We have used different classifiers in WEKA; **(**Bayesian Logistic Regression, Multilayer Perceptron, SMO/SVM, Attribute Selected Classifier, ClassificationViaRegression, NBTree, J48 and Random Forest**).**

C. Methods have been used for attribute selection

- CfsSubset Eval used as an attribute evaluator along with search method “Best First”
- CfsSubset Eval used as an attribute evaluator along with search method “Greedy Stepwise”.

First we used decision tree and ID3/ C4.5 algorithms or the selection of attributes on the bases of high information gain. To get accurate results we used two methods one by one to select attributes that are written above. Same attributes have been selected by both methods we have used for the selection of attributes.

We also did ranking to check that which attributes are more importan. For this purpose, we used search

method “Ranker”:

- InfoGainAttribute Eval used as an attribute evaluator along with by search method “Ranker”.

TABLE-2: The ranking list of attributes

Ranking	Attributes
0.10279	TB
0.09101	ALB
0.08626	DB
0.06656	A/G
0.06578	TP
0.02982	Alk Phos
0.02114	Age
0.02048	Sgot
0.00477	Gender
0	Sgpt

After doing ranking of attributes are in order (TB 0.10279, ALB 0.09101, DB 0.08626, A/G 0.06656, TP 0.06578, Alk Phos 0.02982, Age 0.02114, Sgot 0.02048, Gender 0.00477 and Spgt 0).

B. Evaluation Methodology. By using attribute subset selection we have prepared two types of datasets from Indian liver disease patients dataset to calculate results. So that we can analyze, which classifiers give more accurate results for the prediction of liver disease. Two ways that used to predict liver disease are following:

1. Whole Attribute Dataset

2. Selected Attribute Dataset

a) Whole Attribute Dataset. In this method, we have used dataset with **11** different attributes. Attributes are used in it; **A**ge of the patient, **G**ender of the patient, **T**otal Bilirubin, **D**irect Bilirubin,**A**lkaline Phosphatase, **A**lanine Aminotransferase, **A**spartate Aminotransferase, **T**otal Proteins, **A**lbumin, **A**lbumin and Globulin Ratio, **C**lass label (is-patient): field used to split the data into two sets (patient with liver disease, or with no disease)

b) Selected Attribute Dataset (Proposed Methodology). For selecting attributes, we used Attribute Evaluator along with “search methods”. Decision tree and ID3/C4.5 used to find high information gain attribute. The selected attributes are**: T**otal Bilirubin, **D**irect Bilirubin, **T**otal Proteins, **A**lbumin, **A**lbumin and Globulin Ratio and **C**lass label (is-patient); field used to split the data into two sets (patient with liver disease, or no disease).

C. Evaluation Measures. Two types of evaluation measures have used that help to find best classifier:

- Accuracy
- Computation Time

a) Results and Analysis. For our experiments we have used Indian Liver Disease Dataset that have been taken from Kaggle open source database repository to predict which patients have liver disease and which are not. For this purpose, we have used different data mining classifiers to calculate their accuracies and computation times that which classifier give better accuracy as compared to others on two types of datasets that have been prepared by us.

b) Whole Attribute Dataset. Classifiers have applied to the whole Indian liver disease dataset in which 11 attributes are included along with Class attribute. In results, different classifiers show different accuracies with computation time that are given below in Table III. In this case Bayesian Logistic

Regression algorithm gives higher accuracy with less computation time as compared to others.

TABLE-3: Classifiers accuracy and computation time for Whole Attribute Dataset

Classifier	Accuracy (%)	Computational time (Sec)
Bayesian Logistic Regression	71.3551%	0.02Sec
Multilayer Perceptron	68.7822%	1.13Sec
SMO/SVM	71.3551%	0.05Sec
AttributeSelected Classifier	67.4099%	0.13Sec
ClassificationViaRegression	70.4974%	0.87Sec
NBTree	67.2384%	0.63Sec
J48	68.4391%	0.03Sec
Random Forest	69.9828%	0.61Sec

c) Selected Attribute Dataset. Classifiers Applied on selected attributes dataset in which TB, DB, TP, ALB, and A/G are included along with the class attribute. This is our proposed methodology which gives more accurate results with the less number of attributes. It also helps the system to work more efficiently with less computation time and early decision making. The results are following:

TABLE-4:. Classifiers accuracy and computation time for Selected Attribute Dataset

Classifier	Accuracy (%)	Computational time (Sec)
Bayesian Logistic Regression	71.3551%	0.02Sec
Multilayer Perceptron	70.8405%	0.67Sec
SMO/SVM	71.3551%	0.01Sec
Attribute Selected Classifier	68.6106%	0.03Sec
ClassificationViaRegression	71.012%	0.13Sec
NBTree	68.6106%	0.12Sec
J48	68.6106%	0.02Sec
Random Forest	72.0412%	0.33Sec

d) *Graphical representation of the Classifiers comparison*

Different classifiers have been used on two different datasets that are; whole attribute dataset and selected attribute dataset [11, 12]. Both datasets give different results of classifiers with evaluation measures

(accuracy and computation). As you can see here in graphical representation given below whole attribute dataset give highr accuracy with Bayesian logistic regression and selected attribute dataset gives higher accuracy with Random Forest.

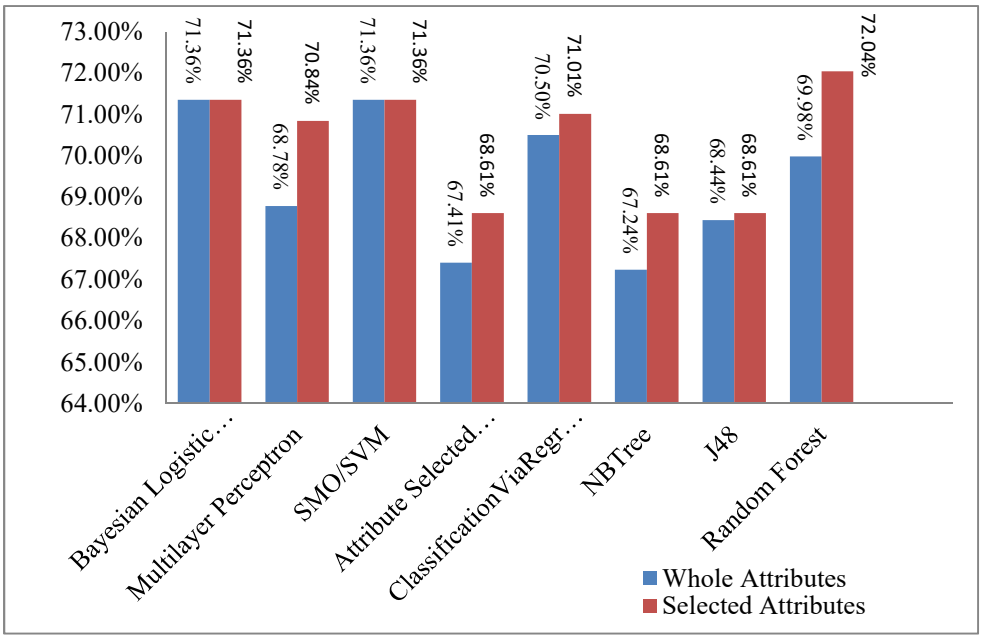

Figure-3.0: Classifiers Comparison *(accuracy and computation time)*

e) **Achieved Accuracies of Classifiers**

TABLE-5: Achieved accuracies of algorithms

Liver Detection Method	**Classifier**	**Accuracy (%)**	**Computation Time (Sec)**
Whole Attribute Dataset Results:			
	Bayesian Logistic Regression	71.3551%	0.02 Sec
Selected Attribute Dataset Results:			
	Random Forest	72.0412%	0.33Sec

As you can see here our proposed methodology gives more accurate results by using "Random forest" classifier with less computation time as compared to other classifiers. The highest accuracy is **72.0412%** with selected attributes, whereas the highest accuracy achieved with all attributes is **71.3551%** through Bayesian Logistic Regression. So from these results we analyzed that we can predict liver disease more accurately with less number of attributes. It provides ease to get more accurate results as well as improved system including these parameters i.e., the efficiency of the system can be increased, early decision making, less time and space required.

Conclusion. It's been very difficult in the recent years to predict patients liver disease in the early stages. Doctors required so many data of patients to predict the liver disease. Proposed methodology provides help to predict the liver disease by acquiring less data. In this paper, different data mining classification techniques are used in which Bayesian Logistic Regression, Multilayer Perceptron, SVM, Attribute Selected Classifier, Classification via Regression, NBTree, J48 and Random Forest is involved to evaluate the liver disease in patients. We have made two types of datasets i.e., Whole Attributes Dataset (WAD) and Selected Attributes Dataset (SAD), and experimented different data mining classifiers to predict liver disease. From our experiments we concludes that SAD provide more accurate results with less computation time by using "Random Forest" as compared to all other classifiers. It also provides improved system including these parameters i.e., the efficiency of the system can be increased, early decision making, less time and space required.

Future work. For future study, the CT scans data of the abdomen can be used to achieve more accurate results in the prediction of liver disease at an early stage. For this purpose, we will use a pixel segmentation technique for real time images of liver disease patients. Firstly, we will find LFT dataset results and then compare and match with the results of CT scans images data. Through this approach there will be more chances to achieve high accuracy for prediction.

REFERENCES

[1] Karthik, S., Priyadarishini, A., Anuradha, J., & Tripathy, B. K. (2011). Classification and rule extraction using rough set for diagnosis of liver disease and its types. *Adv Appl Sci Res*, *2*(3), 334-345.

[2] Sindhuja, D., & Priyadarsini, R. J. (2016). A survey on classification techniques in data mining for analyzing liver disease disorder. *International Journal of Computer Science and Mobile Computing*, *5*(5), 483-488.

[3] KIRUBA, H. R., & ARASU, G. T. (2014). AN INTELLIGENT--AGENT BASED FRAMEWORK FOR LIVER DISORDER DIAGNOSIS USING ARTIFICIAL INTELLIGENCE TECHNIQUES. *Journal of Theoretical & Applied Information Technology*, *69*(1).

[4] Saranya, A., & Seenuvasan, G. (2017). A COMPARATIVE STUDY OF DIAGNOSING LIVER DISORDER DISEASE USING CLASSIFICATION ALGORITHM.

[5] Aneeshkumar, A. S., & Venkateswaran, C. J. (2012). Estimating the surveillance of liver disorder using classification algorithms. *International Journal of Computer Applications*, *57*(6).

[6] Dhamodharan, S. (2014, May). Liver disease prediction using bayesian classification. In *4th National Conference on Advanced Computing, Applications & Technologies* (pp. 1-3).

[7] Rajeswari, P., & Reena, G. S. (2010). Analysis of liver disorder using data mining algorithm. *Global journal of computer science and technology*.

[8] Banu, MA Nishara, and B. Gomathy. "Disease Predicting System Using Data Mining Techniques". International Journal of Technical Research and Applications 1.5 (2013): 41-45.

[9] Alfisahrin, S. N. N., & Mantoro, T. (2013, December). Data mining techniques for optimization of liver disease classification. In *2013 International Conference on Advanced Computer Science Applications and Technologies (ACSAT)*(pp. 379-384). IEEE.

[10] Cheung, N. (2001). Machine learning techniques for medical analysis. *School of Information Technology and Electrical Engineering*.

[11] Ehsan, A., Mahmood, K., Khan, Y. D., Khan, S. A., & Chou, K. C. (2018). A novel modeling in mathematical biology for classification of signal peptides. *Scientific reports*, *8*(1), 1039.

[12] Butt, A. H., Rasool, N., & Khan, Y. D. (2017). A treatise to computational approaches towards prediction of membrane protein and its subtypes. *The Journal of membrane biology*, *250*(1), 55-76.